Tema 3. Relación entre dos variables cuantitativas

Resumen del tema

3.1. Diagrama de dispersión

Cuando sobre cada individuo de una población se observan simultáneamente dos características cuantitativas X e Y, se dice que se está observando una *variable estadística bidimensional*, que se representa por (X, Y).

La representación gráfica más usual es el **diagrama de dispersión** o **nube de puntos**, que consiste en situar en un sistema de ejes coordenados los puntos que resultan de tomar en el eje horizontal los valores de una de las variables y en el eje vertical los valores de la otra.

3.2. Coeficiente de correlación lineal

• **Covarianza** entre X e Y:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n} = \frac{\sum_{i=1}^{n} x_i\, y_i}{n} \; - \; \overline{x}\;\overline{y}\,.$$

De la fórmula anterior se deduce que la unidad de medida de s_{xy} es el producto de la unidad de X por la unidad de Y.

• **Coeficiente de correlación lineal** de Pearson entre X e Y:

$$r_{xy} = \frac{s_{xy}}{s_x\, s_y}\,.$$

• De la fórmula anterior se deduce que r_{xy} no tiene unidad de medida.

• *Propiedad del coeficiente de correlación lineal*: el resultado de r_{xy} siempre está comprendido entre -1 y 1; es decir,

$$-1 \leq r_{xy} \leq 1\,.$$

• *Interpretación descriptiva del coeficiente de correlación lineal*:

⋆ Si $r_{xy} > 0$, existe relación lineal directa entre X e Y; es decir, al aumentar la variable X, aumenta la variable Y.

⋆ Si $r_{xy} < 0$, existe relación lineal inversa entre X e Y; es decir, al aumentar la variable X, disminuye la variable Y.

- ⋆ Si $r_{xy} = 1$, existe dependencia lineal directa exacta entre X e Y; es decir, los puntos del diagrama de dispersión están situados sobre una línea recta de pendiente positiva.
- ⋆ Si $r_{xy} = -1$, existe dependencia lineal inversa exacta entre X e Y; es decir, los puntos del diagrama de dispersión están situados sobre una línea recta de pendiente negativa.
- ⋆ Si $r_{xy} = 0$, no existe dependencia lineal entre X e Y.
- ⋆ Cuanto más se aproxime r_{xy} a -1 o a 1, más dependencia lineal existe entre X e Y. Y cuanto más se aproxime r_{xy} a 0, más independencia lineal existe entre X e Y.

3.3. Recta de regresión

• **Recta de regresión de Y sobre X**: aquella que permite predecir los resultados de la variable Y a partir de los valores de la variable X.

• Ecuación de la recta de regresión (mínimo cuadrática) de Y sobre X:

$$\hat{Y} = A + B\,X\,,$$

donde:

$$B = \frac{s_{xy}}{s_x^2} = r_{xy}\,\frac{s_y}{s_x}\,,$$

$$A = \overline{y} - B\,\overline{x}\,.$$

• **Recta de regresión de X sobre Y**: aquella que permite predecir los resultados de la variable X a partir de los valores de la variable Y.

• Ecuación de la recta de regresión (mínimo cuadrática) de X sobre Y:

$$\hat{X} = A^* + B^*\,Y\,,$$

donde:

$$B^* = \frac{s_{xy}}{s_y^2} = r_{xy}\,\frac{s_x}{s_y}\,,$$

$$A^* = \overline{x} - B^*\,\overline{y}\,.$$

Ejemplos que se van a resolver en clase

Ejemplo 3.1. La tabla siguiente muestra la vejez (años desde su publicación) y la frecuencia de uso (número de veces que se consulta en un año) de ocho libros:

Tabla 3.1

Vejez del libro	1	3	2	4	3	5	4	3
Frecuencia de uso	40	18	30	21	26	10	13	35

Dibujar el diagrama de dispersión.

Ejemplo 3.2. Con los datos de la Tabla 3.1 calcular el coeficiente de correlación lineal entre ambas variables. ¿Cómo se puede calificar el grado de relación lineal: muy fuerte, fuerte, moderado, débil o muy débil? ¿La relación es directa o inversa? Razonar las respuestas.

Ejemplo 3.3. Con los datos de la Tabla 3.1 determinar la ecuación de la recta de regresión de la frecuencia de uso sobre la vejez del libro. Sobre el mismo gráfico en el que se ha hecho el diagrama de dispersión, representar gráficamente la recta de regresión. Estimar el número anual de veces que se prestaría un libro publicado hace 6 años. ¿Es fiable esta estimación? Justificar la respuesta.

Ejemplo 3.4. Con los datos de la Tabla 3.1 determinar la ecuación de la recta de regresión de la vejez del libro sobre la frecuencia de uso. Predecir la vejez de un libro que no fuese consultado ninguna vez durante todo el año. ¿Es fiable esta predicción? ¿Por qué?

Problemas propuestos

Problema 3.1. El número de libros prestados a los estudiantes y a los profesores de los diferentes departamentos de una universidad en un curso académico determinado ha sido:

Departamento	Estudiantes	Profesores
Agricultura	396	70
Antropología	1.122	340
Biología	311	273
Botánica	562	181
Cristalografía	149	33
Física	1.446	704
Geología	1.579	556
Informática	557	233
Ingeniería	1.044	434
Matemáticas	710	437
Mineralogía	52	22
Psicología	1.153	495
Química	737	473
Zoología	1.343	462

a) Dibujar el diagrama de dispersión.

b) Calcular el coeficiente de correlación lineal entre ambas variables. ¿Cómo se puede calificar el grado de relación lineal entre ambas variables: muy fuerte, fuerte, moderado, débil o muy débil? Razonar la respuesta.

c) Determinar la ecuación de la recta de regresión del número de libros prestados a los estudiantes sobre el número de libros prestados a los profesores. Estimar el número de libros prestados a los estudiantes que puede esperarse cuando el número de libros prestados a los profesores sea de 400. ¿Es fiable esta estimación? Justificar la respuesta.

Problema 3.2. El tamaño de la población y el número de libros prestados por las bibliotecas de once ciudades fue:

Población $\times$ 100.000	Nº de préstamos $\times$ 100.000
$114'5$	$86'0$
$25'9$	$35'8$
$4'2$	$51'3$
$7'5$	$47'3$
$6'7$	$7'5$
$6'5$	$94'7$
$6'0$	$77'0$
$5'9$	$39'9$
$4'6$	$18'0$
$4'5$	$36'0$
$4'3$	$68'9$

a) Calcular el coeficiente de correlación lineal entre ambas variables. ¿Cómo se puede calificar el grado de relación lineal entre ambas variables: muy fuerte, fuerte, moderado, débil o muy débil? Razonar la respuesta.

b) Pronosticar el número de libros prestados por las bibliotecas de una ciudad de un millón de habitantes. Decir si es fiable este pronóstico, razonando la respuesta.

Problema 3.3. Los siguientes datos se refieren al número de libros y de revistas que reciben mensualmente doce bibliotecas elegidas al azar.

libros	revistas
1.090	24
7.420	92
4.200	67
8.250	158
8.810	81
1.620	59
3.840	54
9.400	171
3.630	100
14.100	276
2.500	122
11.470	200

a) Calcular el coeficiente de correlación lineal entre ambas variables. ¿Cómo se puede calificar el grado de relación lineal entre ambas variables: muy fuerte, fuerte, moderado, débil o muy débil? Razonar la respuesta.

b) Estimar el número de revistas que recibiría una biblioteca en un mes en el que le enviaran 5.000 libros. ¿Es fiable esta estimación? Justificar la respuesta.

Soluciones de los problemas propuestos

Solución del problema 3.1. Sea X =*número de libros prestados a los estudiantes de cada departamento de la determinada universidad, durante el determinado curso académico* e Y =*número de libros prestados a los profesores de cada departamento de la determinada universidad, durante el determinado curso académico.*

(a) El diagrama de dispersión o nube de puntos consiste en situar en un sistema de ejes coordenados los puntos que resultan de tomar en el eje horizontal los valores de una de las variables y en el eje vertical los valores de la otra.

(b) El coeficiente de correlación lineal entre X e Y es $r_{xy} = 0'8851$. Como este coeficiente está bastante próximo a 1, la relación lineal entre ambas variables se puede calificar de fuerte.

(c) La recta de regresión del número de libros prestados a los estudiantes sobre el número de libros prestados a los profesores es la recta de regresión de X sobre Y, cuya ecuación es: $\widehat{X} = 95'9530 + 2'0831\, Y$

El número de libros prestados a los estudiantes que puede esperarse cuando el número de libros prestados a los profesores sea de 400 es: $\widehat{X} = 95'9530 + 2'0831 \cdot 400 = 929'193$; es decir, 929 libros, aproximadamente.

Esta estimación es bastante fiable ya que el coeficiente de correlación lineal está bastante próximo a 1 y, por tanto, los puntos de la recta de regresión y los puntos del diagrama de dispersión están bastante próximos.

Solución del problema 3.2. Sea X =*número de habitantes de cada ciudad, multiplicado por 100.000* e Y =*número de libros prestados por la biblioteca de cada ciudad, multiplicado por 100.000.*

(a) El coeficiente de correlación lineal entre X e Y es $r_{xy} = 0'3846$. Como este coeficiente está próximo a cero, la relación lineal entre ambas variables se puede calificar de débil.

(b) Para hacer este pronóstico hay que determinar la ecuación de la recta de regresión de Y sobre X, que es: $\widehat{Y} = 45'4902304 + 0'32532773\, X$.

El pronóstico del número de libros prestados por las bibliotecas de una ciudad de un millón de habitantes es: $\widehat{Y} = 45'4902304 + 0'32532773 \cdot 10 = 48'7435077$ multiplicado por 100.000=4.874.350'77 libros; es decir, aproximadamente 4.874.351 libros.

Este pronóstico es poco fiable ya que el valor del coeficiente de correlación lineal entre X e Y está próximo a cero y, por tanto, los puntos de la recta de regresión y los puntos del diagrama de dispersión están bastante alejados.

Solución del problema 3.3. Sea X =*número de libros recibidos mensualmente por cada biblioteca* e Y =*número de revistas recibidas mensualmente por cada biblioteca.*

(a) El coeficiente de correlación lineal entre X e Y es $r_{xy} = 0'8605$. Como este coeficiente está bastante próximo a 1, la relación lineal entre ambas variables se puede calificar de fuerte.

(b) Para hacer esta estimación hay que determinar la recta de regresión de Y sobre X, que es: $\widehat{Y} = 21'6844 + 0'0150\, X$.

La estimación del número de revistas que recibiría una biblioteca en un mes en el que le enviaran 5 000 libros es: $\widehat{Y} = 21'6844 + 0'0150 \cdot 5\,000 = 96'6082$; es decir, 97 libros, aproximadamente.

Esta predicción es bastante fiable ya que el valor del coeficiente de correlación lineal entre X e Y está bastante próximo a 1 y, por tanto, los puntos de la recta de regresión y los puntos del diagrama de dispersión están bastante próximos.